文章编号：1672-3961(2024)06-0001-07 DOI：10.6040/j.issn.1672-3961.0.2023.157

DMK*K*-means——一种深度多核 *K*-means 聚类算法

王梅[1,2]，宋凯文[1]，刘勇[3,4*]，王志宝[1]，万达[1]
（1.东北石油大学计算机与信息技术学院，黑龙江 大庆 163318；2.黑龙江省石油大数据与智能分析重点实验室，黑龙江 大庆 163318；3.中国人民大学高瓴人工智能学院，北京 100049；4.大数据管理与分析方法研究北京市重点实验室（中国人民大学信息学院），北京 100049）

摘要：针对传统 *K*-means 的聚类效果容易受到样本分布影响，且核函数表示能力不强导致对于复杂问题的聚类效果表现不佳的问题，利用深度核的强表示性并通过多核集成方式，提出一种具有强表示能力且分布鲁棒的深度多核 *K*-means（deep multiple kernel *K*-means，DMK*K*-means）聚类算法。构建具有强表示能力的深度多核网络架构，在新的特征空间进行 *K*-means 聚类；基于 Kullback-Leibler（KL）散度的聚类损失函数衡量该算法与 2 种基准聚类方法的差异；将该聚类算法建模成高效的端到端学习问题，利用随机梯度下降算法更新优化深度多核网络的权重参数。在多个标准数据集上进行试验，结果表明，相比于 *K*-means、径向基函数核 *K*-means（radial basis function kernel *K*-means，RBFKKM）及其他多核 *K*-means 聚类算法，该算法在聚类精度、归一化互信息和调整兰德系数指标上均有明显提升，验证该算法的可行性与有效性。

关键词：*K*-means；核聚类；深度多核学习；数据挖掘；梯度下降
中图分类号：TP391 **文献标志码**：A
引用格式：王梅，宋凯文，刘勇，等. DMK*K*-means——一种深度多核 *K*-means 聚类算法[J].山东大学学报（工学版），2024，54(6)：1-7.
WANG Mei，SONG Kaiwen，LIU Yong，et al. DMK*K*-means：a deep multiple kernel *K*-means clustering algorithm[J]. Journal of Shandong University（Engineering Science），2024，54(6)：1-7.

DMK*K*-means：a deep multiple kernel *K*-means clustering algorithm

WANG Mei[1,2]，SONG Kaiwen[1]，LIU Yong[3,4*]，WANG Zhibao[1]，WAN Da[1]
（1. School of Computer and Information Technology，Northeast Petroleum University，Daqing 163318，Heilongjiang，China；
2. Heilongjiang Key Laboratory of Petroleum Big Data and Intelligent Analysis，Daqing 163318，Heilongjiang，China；3. Gaoling School of Artificial Intelligence，Renmin University of China，Beijing 100049，China；4. Beijing Key Laboratory of Big Data Management and Analysis Method（School of Information，Renmin University of China），Beijing 100049，China）

Abstract：The proposed algorithm，deep multiple kernel *K*-means（DMK*K*-means），addressed the limitations of traditional *K*-means clustering，which was sensitive to sample distribution and exhibited suboptimal performance for complex problems due to its limited expressive power of kernel representations. By leveraging the strong representational capability of deep kernels and employing a multi-kernel ensemble approach，DMK*K*-means constructed a highly expressive deep multiple kernel network architecture and performed *K*-means clustering in a new feature space. The dissimilarity between this algorithm and two baseline clustering methods was quantified using a clustering loss function based on Kullback-Leibler（KL）divergence. The clustering algorithm was modeled as an efficient end-to-end learning problem，and the weight parameters of the deep multiple kernel network were optimized through stochastic gradient descent. Experimental results on multiple standard datasets demonstrated the superiority of the proposed algorithm over *K*-means，radial basis function kernel *K*-means（RBFKKM），and other multi-kernel *K*-means clustering algorithms in terms of clustering accuracy，normalized mutual information，and adjusted rand index. These findings validated the feasibility and effectiveness

收稿日期：2023-07-03
基金项目：国家自然科学基金资助项目（51774090，62076234）；黑龙江省博士后科研启动金资助项目（LBH-Q20080）；黑龙江省自然科学基金资助项目（LH2020F003）；黑龙江省高校基本科研业务费资助项目（KYCXTD201903，YYYZX202105）
第一作者简介：王梅（1976— ），女，河北保定人，教授，硕士生导师，博士，主要研究方向为机器学习、模型选择、核方法。
E-mail：wangmei@nepu.edu.cn
*通信作者简介：刘勇（1986— ），男，湖南益阳人，副研究员，博士生导师，博士，主要研究方向为大规模机器学习及统计机器学习理论。
E-mail：liuyonggsai@ruc.edu.cn

of the proposed algorithm.

Keywords: *K*-means; kernel clustering; deep multiple kernel learning; data mining; gradient descent

0 引言

聚类算法根据数据之间的相似性或距离等关系,将数据划分成若干个不相交的子集,使簇内相似度高、簇间相似度低[1-3],广泛用于数据挖掘、图像分割和模式识别等领域[4-6]。*K*-means 聚类算法是流行的聚类算法之一,通过迭代求解使聚类损失最小。然而,由于数据样本分布的多样化,*K*-means 聚类算法在处理非线性可分或高维数据时无法表现出较好的效果。

为克服这些限制,学者们在 *K*-means 聚类中引入核方法,可以将样本从原始空间映射到新的特征空间,并由此提出核 *K*-means 聚类算法[7]。该类算法能够将数据特征映射到高维空间,进行标准的 *K*-means聚类,由此可以处理在原始空间线性不可分的数据。然而这些工作大部分是基于单个核函数的核聚类方法,鲁棒性较弱,且针对数据分布的不同需要选择不同的核函数,使核函数的选择成为影响聚类性能的关键因素。为解决单核聚类存在的局限性,研究者们逐渐将多核聚类引入研究并加以应用[8-13]。这些方法通过组合多个核函数可以捕捉数据的更多特征和结构,根据数据特点的不同选择不同的核函数,但多核聚类通常采用交替优化基核参数和聚类划分矩阵的方法进行求解[14],容易使目标函数陷入局部最优解。

近年来,多层次结构的深度模型在深度学习领域受到广泛研究与应用,深度神经网络通过对输入数据进行多层非线性处理构造新的特征[15],为各种下游任务提供强大的支持,给对非线性数据进行聚类的任务带来了新思路。

与多层次深度模型相比,无论基于单核还是多核的核聚类算法,其结构形式都是浅层次的,数据表示能力仍存在局限性,对非线性数据建模能力有限,且对输入数据的敏感度较高,容易受异常样本干扰。为此,本研究提出一种深度多核 *K*-means (deep multiple kernel *K*-means, DMK *K*-means)聚类算法。每对数据样本经过具有强表示能力的深度多核网络模型进行映射,经过多层次的多核迭代方式学习新的、鲁棒性较强的特征表示;基于学习到的核函数在新的特征上进行 *K*-means聚类;基于 Kullback-Leibler(KL)散度损失函数,通过联合训练,利用反向传播和随机梯度下降算法优化网络参数和*K*-means聚类结果。

1 相关工作

1.1 核聚类

核聚类可以将原始空间线性不可分的数据映射到高维特征空间,在新的特征空间中形成清晰的决策边界完成聚类工作。为了理解并指导核聚类的发展,学者们进行了大量的研究[16-19]。文献[20]提出一种核聚类方法,将样本从输入空间映射到高维特征空间进行聚类,相比经典聚类方法有较大的性能改进;针对 *K*-means 聚类算法容易受聚类中心初始位置选择影响的问题,文献[21]提出一种全局核 *K*-means 聚类算法,是一种确定性和增量式的核聚类算法;文献[22]为核*K*-means的过度聚类风险提供了近乎最优的界限,提高了风险界限。但基于单核的聚类算法容易受所选择的核函数的影响,在实际应用中,往往需要提取多组特征得到互补的信息表示样本,从而得到更好的聚类效果,因此学者们对多核聚类开展了大量研究工作。

现有的部分多核聚类算法一般假定基核的线性组合为最优核,例如:文献[23]提出一种多核模糊 *C*-means 聚类算法,通过合并多个核自动调整权值,使聚类效果不再受核选择的较大影响;文献[24]处理多个数据源进行聚类,提出交替极小化聚类隶属度和基核系数的一种核 *K*-means 优化算法。但上述通过优化核函数组合系数的方法,聚类时容易忽略核补集中鲁棒性更强的核函数。另有部分多核聚类算法通过低秩优化学习共识矩阵,例如:文献[25]通过多核构造的转移概率矩阵恢复一个共享的低秩矩阵,让其作为标准马尔可夫链方法聚类的输入;文献[26]从多视图角度提取、挖掘样本特征信息,将核矩阵和谱聚类优化结合到一个框架中学习低秩矩阵。然而,无论基于单核还是多核的核聚类算法,其表现形式均为浅层次(即单层次)结构,导致核函数的表示能力不强。为此,利用深度核的强表示性,本研究提出深度多核 *K*-means 聚类算法。

1.2 深度多核学习

深度学习目前是机器学习中热门的研究领域,

一般深度学习模型都具有多层层次结构,往往比浅层结构具有更好的表现效果。因此,应用深度学习思想,通过将多核学习与深度多层次结构相结合,学者们开展了大量的研究工作。与基于平均权重系数的简单多核学习方法相比,现有改进多核学习的方法往往没有表现出更好的性能。为了改进多核学习的体系结构和方法,得到多层结构中多个核函数的最优组合,学者们对这项工作开展了研究:文献[27]将核函数与深度架构相结合,提出一种名为反余弦核的核函数,通过调整核参数迭代重复试验,得到比普通多核学习方法更好的试验结果,但当使用其他特定核函数时,取得的效果并不理想。为解决上述问题,文献[26]提出双层结构的多核学习方法,依据多核学习思想,将多个核函数组合为整体结构的第 1 层,且每个核函数相对应关联一个权值,第 2 层结构由单个高斯核构成,但该方法无法优化 2 层以外的网络结构。为增加网络层次,文献[28]通过调整文献[29]的估计误差,成功优化了多层次的深度多核学习算法,证明了某些条件下深度多核的泛化误差上限小于深度前馈网络,但并没有用多核学习算法与该算法对比,同时超过 2 层的深度结构并没有表现出更好的试验结果。针对多层网络结构优化问题,文献[30]受训练神经网络的启发,进一步提出使用自适应反向传播算法优化网络。在基于深度多核学习应用方面,文献[31]成功使用深度多核网络模型进行图像标注,并在数据集上取得了良好的结果。文献[32]结合深度多核学习解决了合成孔径雷达图像目标识别问题。

2　深度多核 *K*-means 聚类

2.1　核 *K*-means 聚类

假设给定一组有限样本集合 $\boldsymbol{T}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\boldsymbol{x}_i\in\mathbf{R}^m$,其中 n 为样本数,$\boldsymbol{x}_i$ 为第 i 个样本,每个样本由一个特征向量表示,m 为特征向量的维度。目的是将 n 个样本划分到 k 个簇中。初始设定 k 个聚类中心 $\boldsymbol{c}_1$、$\boldsymbol{c}_2$、…、$\boldsymbol{c}_k$,聚类中心的更新可表示为 $\boldsymbol{c}_j=\frac{1}{N_j}\sum_{i=1}^{N_j}\boldsymbol{x}_i$, $j=1,2,\cdots,k$,其中 N_j 为第 j 类中的样本数,*K*-means 聚类基于最小误差平方和准则,其目标函数

$$J=\arg\min_{\boldsymbol{c}_j}\sum_{j=1}^{k}\sum_{\boldsymbol{x}_i\in\boldsymbol{c}_j}\|\boldsymbol{x}_i-\boldsymbol{c}_j\|^2。\tag{1}$$

核 *K*-means 是标准 *K*-means 聚类算法的推广,核 *K*-means 将原始空间数据通过非线性变换映射到高维空间,并在新的特征空间进行线性划分。设存在一个非线性映射 $\varphi:\boldsymbol{x}_i\in\boldsymbol{X}\rightarrow\boldsymbol{H}$,将 $\boldsymbol{x}_i$ 映射到一个可再生核希尔伯特空间 $\boldsymbol{H}$,核 *K*-means 试图最小化特征空间中的聚类误差,则核 *K*-means 聚类的目标函数

$$J_{\text{kernel}}=\arg\min_{\boldsymbol{c}_j}\sum_{j=1}^{k}\sum_{\boldsymbol{x}_i\in\boldsymbol{c}_j}\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}_j^{\varphi}\|^2,\tag{2}$$

式中,$\boldsymbol{c}_j^{\varphi}$ 为新特征空间中第 j 类聚类中心,$\varphi(\boldsymbol{x}_i)$ 为样本 i 在新特征空间的映射。

2.2　深度多核网络

本研究中深度多核网络架构是对文献[16]中工作的扩展。将文献[16]提出的两层多核学习方法升级为多层,使用深度架构重新定义多核,学习到的全局核为激活函数的多层线性组合,每个激活函数包含多个基核函数或中间函数在多个特征上的组合,数据经过深层次多核迭代映射,能够明显划分样本特征间的差异。深度多核网络与下游聚类任务结合,进行联合优化。

以 3 层深度多核网络架构为例,其中第 1 层为输入层,数据通过第 1 层的多个基核传入深度多核网络中,其中基础核函数集合包括径向基函数(radial basis function, RBF)核和 sigmoid 核;中间几层为隐藏层,此层的每个核单元是对前一层的所有核值进行组合,再经过激活函数进行非线性转换;最后一层为输出层,倒数第 2 层的核单元重复之前的过程,最终深度多核网络输出提取到的特征核值。递归前馈过程在前一层核值的线性组合上计算一个非线性激活函数,直到输出。这种递归形式定义如下:

$$\boldsymbol{K}_p^{(l)}(\cdot,\cdot)=g\left(\sum_q\boldsymbol{w}_{p,q}^{(l-1)}\boldsymbol{K}_q^{(l-1)}(\cdot,\cdot)\right),\ l=2,\cdots,L,\tag{3}$$

式中,$\boldsymbol{K}_q^{(l-1)}(\cdot,\cdot)$ 为第 $l-1$ 层的第 q 个位置处的核值,$\boldsymbol{w}_{p,q}^{(l-1)}$ 为连接第 l 层和第 $l-1$ 层的核单元权值,L 为网络的层数,$g(\cdot)$ 为深度多核网络中的非线性激活函数(如双曲函数或指数函数)。双曲函数可使学习到的数值更加稳定,同时保证输出的核是半正定的。

由以上分析可知,深度多核网络具有类似神经网络的深度多层次结构,与一般的多核学习方法或浅层次的核组合结构相比,当处理较为复杂的数据样本时效果更加突出,深度多核网络对输入数据进行多层次的非线性处理,学习到有利于下游任务的特征表示。

2.3 DMK*K*-means 聚类算法

2.3.1 聚类损失

本研究提出一种深度多核聚类算法,能够根据设定好的聚类损失函数确定最优权重。与传统神经网络模型不同,深度多核网络的输出为一个核函数,即某些隐含的再生希尔伯特空间中 2 个输入示例的内积。因此,本研究提出一个基于 KL 散度的聚类损失函数,衡量模型输出与目标之间的差异。通过基于单核或欧氏距离的 *K*-means 聚类方法对数据进行聚类,根据聚类结果计算 KL 散度。现定义 DMK*K*-means 聚类算法损失函数

$$D=\mathrm{KL}(P\parallel Q)=\sum_i\sum_j p_{ij}\ln\frac{p_{ij}}{q_{ij}},\tag{4}$$

式中:P 和 Q 分别为目标分布和软标签分布;q_{ij} 为通过学生 t-分布衡量样本 $\boldsymbol{x}_i$ 属于第 j 类的概率,$q_{ij}=\dfrac{(1+\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel^2)^{-1}}{\sum\limits_j(1+\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel^2)^{-1}}$;$p_{ij}$ 为样本 $\boldsymbol{x}_i$ 属于第 j 类的真实概率,$p_{ij}=\dfrac{q_{ij}^2/\sum\limits_i q_{ij}}{\sum\limits_j(q_{ij}^2/\sum\limits_i q_{ij})}$。

算法的目标是通过输入数据样本对,发现数据之间的特征表示关系,训练深度多核网络学习到一个距离度量函数,即一个至少能与基于欧氏距离或单核 *K*-means 聚类算法相比聚类效果表现更好的核函数。因此,本研究决定使用 KL 散度评估当前深度多核网络学习到的核与 2 个基准方法之间的差异。

2.3.2 优化过程

受深度神经网络训练的启发,本研究采用反向传播和小批量随机梯度下降法最小化损失函数。深度多核网络的输出是输入样本 $\boldsymbol{x}_p$ 和 $\boldsymbol{x}_q$ 在某个特征空间中的距离。使用这种距离度量函数对数据进行 *K*-means 聚类,再计算 KL 散度。q_{ij} 的计算式中,样本到质心的距离可以用单个核函数或深度多核网络的输出代替。D 相对于样本数据 $\boldsymbol{x}_i$ 和聚类中心 $\boldsymbol{c}_i$ 的梯度分别为:

$$\frac{\partial D}{\partial \boldsymbol{x}_i}=2\sum_{j=1}^{k}(1+\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel^2)^{-1}(p_{ij}-q_{ij})(\boldsymbol{x}_i-\boldsymbol{c}_j),\tag{5}$$

$$\frac{\partial D}{\partial \boldsymbol{c}_j}=2\sum_{i=1}^{n}(1+\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel^2)^{-1}(p_{ij}-q_{ij})(\boldsymbol{x}_i-\boldsymbol{c}_i)。\tag{6}$$

假定学习率为 λ,小批量数据的样本数为 s,则聚类中心更新为:

$$\boldsymbol{c}_j=\boldsymbol{c}_j-\frac{\lambda}{s}\sum_{i=1}^{s}\frac{\partial D}{\partial \boldsymbol{c}_j},\tag{7}$$

深度多核网络权重更新为:

$$\boldsymbol{w}=\boldsymbol{w}-\frac{\lambda}{s}\sum_{i=1}^{s}\frac{\partial D}{\partial \boldsymbol{w}}。\tag{8}$$

2.3.3 权重参数初始化

网络连接的权重参数对深度多核网络学习到的核度量函数有重要影响。文献[33]发现无监督的预训练可以极大提高模型的泛化能力。受文献[34]的启发,本研究提出一种网络权值初始化的方法,主要思想是保持归一化标准,同时对权重施加先验分布。已知高斯分布及其扩展广泛应用在深度模型的初始化中,因此,本研究使用高斯过程初始化深度多核网络的权值[35]。模型中所有权值的周期核矩阵

$$\boldsymbol{\kappa}^{(w)}=[k_{ij}]_{n\times n},\tag{9}$$

式中 k_{ij} 为 $n\times n$ 核矩阵中第 i 行第 j 列的元素,$k_{ij}=\exp\left(-\dfrac{2\sin^2(\frac{\parallel \boldsymbol{w}_i-\boldsymbol{w}_j\parallel}{2})}{l^2}\right)$。式(9)可认为是特征空间中权重的先验分布,从中采样可以得到 $\boldsymbol{w}$ 的初始值,有关此方法的详细信息可参考文献[36]。

2.3.4 本研究算法

综上所述,本研究提出的 DMK*K*-means 聚类算法具体流程如算法 1 所示。

算法 1 DMK*K*-means 聚类算法

输入 聚类数 k、样本数据 T、训练样本对 T'、学习率 λ、停止阈值 δ、网络初始权重 $\boldsymbol{w}$。

输出 网络权重参数 $\boldsymbol{w}^*$。

(1) 基于欧氏距离对数据集 $\boldsymbol{T}$ 进行 *K*-means 聚类,求出分布 $Q_1=c_{\mathrm{luster}}(\boldsymbol{T},E_{\mathrm{uclidean}},k)$;

(2) 基于 RBF 核对数据集 $\boldsymbol{T}$ 进行核 *K*-means 聚类,求出分布 $Q_2=c_{\mathrm{luster}}(\boldsymbol{T},R_{\mathrm{BF}},k)$;

(3) **while** true **do**

(4) $D_{\mathrm{ist}}=c_{\mathrm{al}}(K(w),\boldsymbol{T'})$,即计算所有样本对的深度多核网络输出;

(5) $P=c_{\mathrm{luter}}(D,D_{\mathrm{ist}},k)$,即基于深度多核输出的核矩阵进行 *K*-means 聚类;

(6) 根据式(4)计算 $D_1=\mathrm{KL}(P|Q_1)$;

(7) 根据式(4)计算 $D_2=\mathrm{KL}(P|Q_2)$;

(8) 根据式(8)更新 $\boldsymbol{w}=B_{\mathrm{P}}(K(w),D_1,D_2,\lambda)$;

(9) **if** $\Delta(D_1)+\Delta(D_2)<\delta$ **then**;

(10) break while;

(11) **end if**;

(12)　**end while**;

(13) $\boldsymbol{w}^{*}=\boldsymbol{w}$;

(14) 返回 $\boldsymbol{w}^{*}$。

在算法 1 中,$\boldsymbol{T}'$为构建好的一组样本对,作为深度多核网络的输入。对于数据集 $\boldsymbol{T}$,样本对的个数为$|\boldsymbol{T}|\times(|\boldsymbol{T}|-1)/2$。

2.3.5　算法时间复杂度分析

假设数据集规模和样本特征维度分别为 n 和 m,则本研究计算具有 L 层深度多核网络的时间复杂度为 $O(n^2(m+L))$;采用梯度下降法求解优化问题的时间复杂度为 $O(n^2)$;设给定聚类簇数为 k,则 *K*-means 聚类算法的时间复杂度为 $O(knm)$。由此得出,本研究 DMK*K*-means 聚类算法的时间复杂度为 $O(n^2(m+L))+O(n^2)+O(knm)$。

3　试验与分析

3.1　数据集

考虑到数据样本的规模、维度及类别数会影响聚类算法的性能,为验证本研究算法能够在不同数据上有较好表现,在 4 个广泛使用的公开数据集上进行试验。数据集的详细信息如表 1 所示。

表 1　数据集信息

Table 1　Information of datasets

数据集	样本数/个	簇数/个	维度
yeast	1 484	10	8
digits	1 797	10	64
glass	214	9	9
Caltech7	1 474	7	6

3.2　试验方法及结果分析

本研究在算法设计过程中,为了训练深度多核网络学习到一个具有更强表示能力的核函数,选用标准 *K*-means 聚类算法和基于径向基函数核的单核 *K*-means (radial basis function kernel *K*-means, RBFKKM) 聚类算法作为 2 个基准比较方法[20]。同时,又与 3 种多核聚类算法进行了比较,以验证其有效性。其中,平均核 *K*-means (average kernel *K*-means, AKKM) 聚类算法将基础核矩阵进行均匀组合,用于核 *K* 均值的输入;多核 *K*-means (multiple kernel *K*-means, MKKM) 聚类算法对基础核矩阵进行线性组合,并在聚类过程中不断优化权重参数[23];鲁棒多核 *K*-means (robust multiple kernel *K*-means, RMKKM) 聚类算法通过捕捉多核中的噪声结构学习用于聚类的鲁棒低秩核矩阵。

本研究采用 3 种不同的聚类效果评价指标评估 3 种聚类算法的性能,分别是调整兰德系数 I_{AR}、归一化互信息 N_{MI}和聚类精度 A_{CC}。所有聚类算法在每个数据集上分别运行 20 次,统计每次试验结果并累加求平均,3 种评价指标均与聚类效果成正比。

3.2.1　模型分析

根据文献[31-39]在图像标注任务中对节点数目的设置,本研究从 10~20 中选择每层节点的个数。为了确定深度多核网络层数,在 digits 和 glass 数据集上进行先验试验,改变网络结构的层数,记录网络层数 L 对模型性能的影响,通过将层数从 3 改到 10,获取 A_{CC}的变化,试验结果如图 1 所示。

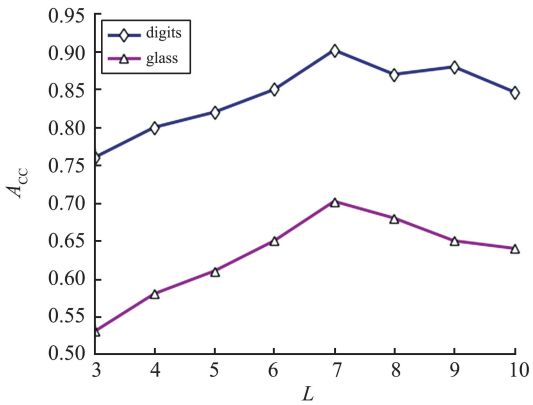

图 1　深度多核网络层数的影响图

Fig.1　The effect of the number of deep multiple kernel network layers

由图 1 可看出:随着网络层数的增加,聚类精度提高,但存在一个最佳层数,超过最佳层数后,模型的准确性下降。选择一个合适的网络层数对试验结果会有所提升,也应找到最优的网络结构平衡模型性能和节约计算成本。鉴于此,本研究将模型层数设置为 7 层。

3.2.2　聚类性能比较

调整兰德系数 I_{AR}是对兰德系数的改进,通常用于衡量 2 个数据分布的吻合程度,I_{AR}的取值范围为 $[-1,1]$,其值越接近 1,代表聚类结果与实际情况越相符。不同算法在各数据集上的 I_{AR}结果如表 2 所示。表 2 中粗体数据表示本研究所提算法在所有数据集上的 I_{AR}均优于其他算法。

表 2　不同算法在 4 个数据集下的 I_{AR}

Table 2　I_{AR} of different algorithms on four datasets

算法	I_{AR}			
	yeast	digits	glass	Caltech7
K-means	0.153 0	0.663 1	0.279 2	0.175 5
RBFKKM	0.154 7	0.675 8	0.264 9	0.183 6
AKKM	0.276 8	0.774 9	0.396 6	0.356 5
MKKM	0.194 6	0.621 2	0.332 1	0.254 8
RMKKM	0.256 3	0.704 0	0.364 7	0.346 9
DMK*K*-means	**0.356 6**	**0.810 1**	**0.451 4**	**0.472 1**

归一化互信息 N_{MI} 是常见的聚类性能评价指标，是指2个事件之间的相关性，常用于衡量数据间的分布吻合程度，取值范围为[0,1]，N_{MI} 越大，代表数据划分到相应类的可能性越大，聚类结果越真实。不同算法在所有数据集上的 N_{MI} 如表3所示。表3中粗体数据表示本研究所提算法在所有数据集上的 N_{MI} 均优于其他算法。

表3 不同算法在4个数据集下的 N_{MI}

Table 3 N_{MI} of different algorithms on four datasets

算法	N_{MI}			
	yeast	digits	glass	Caltech7
K-means	0.271 5	0.732 7	0.425 2	0.475 2
RBFKKM	0.287 4	0.747 0	0.433 6	0.568 1
AKKM	0.317 9	0.801 5	0.546 2	0.658 8
MKKM	0.270 0	0.711 6	0.453 4	0.609 2
RMKKM	0.321 5	0.732 0	0.528 7	0.632 0
DMK*K*-means	**0.373 5**	**0.841 5**	**0.591 0**	**0.687 5**

聚类精度 A_{CC} 是指聚类结果正确的样本数占样本总数的比例，通常用于综合衡量真实标签和聚类标签的匹配程度，其值越接近1，代表聚类效果越好。不同算法在所有数据集上的 A_{CC} 如表4所示。表4中粗体数据表示本研究所提算法在所有数据集上的 A_{CC} 均优于其他算法。

表4 不同算法在4个数据集下的 A_{CC}

Table 4 A_{CC} of different algorithms on four datasets

算法	A_{CC}			
	yeast	digits	glass	Caltech7
K-means	0.417 4	0.781 7	0.545 1	0.197 3
RBFKKM	0.396 3	0.798 6	0.521 8	0.224 9
AKKM	0.453 8	0.885 6	0.632 1	0.338 0
MKKM	0.405 8	0.766 9	0.553 1	0.246 0
RMKKM	0.475 0	0.852 0	0.604 5	0.284 6
DMK*K*-means	**0.526 9**	**0.902 1**	**0.702 2**	**0.376 1**

通过深度多核网络的多层结构和随机梯度下降的优化算法，本研究算法能够更好地捕捉数据的复杂特征和分布，具有更好的表现效果，能够自动学习到更有代表性的特征表示。多核聚类受限于预定义的核函数和迭代优化算法，在表达能力和优化效果上存在局限性，容易陷入局部最优。同时，从以上聚类结果可以看出，本研究算法在所有数据集上均取得了最佳的聚类性能，并且在glass数据集上，相比次优算法，A_{CC} 提高了7%以上，证实了该算法的有效性。

4 结论

本研究基于 *K*-means 和核 *K*-means 聚类方法提出一种新的聚类算法 DMK*K*-means。将数据输入深度多核网络进行特征映射，充分发掘数据之间的相似关系，同时在新的特征空间完成聚类，结合聚类损失，利用反向传播和随机梯度下降法更新网络参数，不断优化网络模型。多个基准数据集上的试验结果验证了本研究算法的有效性。由于深度多核网络的训练计算时间和存储成本随网络层数和基核个数的增加而增大，同时，自动化地选取模型层数也是一个重要问题。在未来的工作中，将考虑采用核近似方法减少算法的时间消耗，并且考虑使用网格搜索等方法，针对样本数据集确定最佳的模型参数。

参考文献：

[1] 章永来，周耀鉴. 聚类算法综述[J]. 计算机应用，2019，39(7)：1869-1882.
ZHANG Yonglai，ZHOU Yaojian. Review of clustering algorithms[J]. Journal of Computer Applications，2019，39(7)：1869-1882.

[2] MADHULATHA T S. An overview on clustering methods[J]. IOSR Journal of Engineering，2012，2(4)：719-725.

[3] XU R，WUNSCH D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks，2005，16(3)：645-678.

[4] 徐金东，赵甜雨，冯国政，等. 基于上下文模糊 *C* 均值聚类的图像分割算法[J]. 电子与信息学报，2021，43(7)：2079-2086.
XU Jindong，ZHAO Tianyu，FENG Guozheng，et al. Image segmentation algorithm based on context fuzzy *C*-means clustering[J]. Journal of Electronics & Information Technology，2021，43(7)：2079-2086.

[5] 姜东明，杨火根. 融合图卷积网络模型的无监督社区检测算法[J]. 计算机工程与应用，2020，56(20)：59-66.
JIANG Dongming，YANG Huogen. Unsupervised community detection algorithm integrating graph convolutional network model[J]. Computer Engineering and Applications，2020，56(20)：59-66.

[6] 刘大莲，田英杰. 可拓数据挖掘在学生成绩分析中的应用研究[J]. 智能系统学报，2022，17(4)：707-713.
LIU Dalian，TIAN Yingjie. Application of extension data mining in student achievement analysis[J]. CAAI Transactions on Intelligent Systems，2022，17(4)：707-713.

[7] SCHÖLKOPF B，SMOLA A，MÜLLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation，1998，10(5)：1299-1319.

[8] ZHAO B，KWOK J T，ZHANG C. Multiple kernel clustering[C]//Proceedings of the 2009 SIAM International Conference on Data Mining. Sparks，USA：

Society for Industrial and Applied Mathematics, 2009: 638-649.

[9] LU Y, WANG L, LU J, et al. Multiple kernel clustering based on centered kernel alignment[J]. Pattern Recognition, 2014, 47(11): 3656-3664.

[10] GÖNEN M, MARGOLIN A A. Localized data fusion for kernel *K*-means clustering with application to cancer biology[J]. Advances in Neural Information Processing Systems, 2014, 27: 1305-1313.

[11] 俞磊, 朱铮, 蒋超, 等. 自适应局部核的最优邻域多核聚类[J]. 控制工程, 2022, 29(1): 182-192.
YU Lei, ZHU Zheng, JIANG Chao, et al. Optimal neighborhood multiple kernel clustering based on adaptive local kernel[J]. Control Engineering of China, 2022, 29(1): 182-192.

[12] JIA L, LI M, ZHANG P, et al. SAR image change detection based on multiple kernel *K*-means clustering with local-neighborhood information[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(6): 856-860.

[13] 欧琦媛, 祝恩. 基于压缩子空间对齐的多核聚类算法[J]. 计算机工程与科学, 2021, 43(10): 1730-1735.
OU Qiyuan, ZHU En. Multiple-kernel clustering based on compressed subspace alignment[J]. Computer Engineering & Science, 2021, 43(10): 1730-1735.

[14] LIU X. Simple MKKM: simple multiple kernel *K*-means [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023:5174-5186.

[15] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[16] ZHUANG J, TSANG I W, HOI S C H. Two-layer multiple kernel learning[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Florida, USA: JMLR, 2011: 909-917.

[17] 王梅, 宋晓晖, 刘勇, 等. 神经正切核 *K*-means 聚类[J]. 计算机应用, 2022, 42(11): 3330-3336.
WANG Mei, SONG Xiaohui, LIU Yong, et al. Neural tangent kernel *K*-means clustering[J]. Journal of Computer Applications, 2022, 42(11): 3330-3336.

[18] BEN-HUR A, HORN D, SIEGELMANN H T, et al. Support vector clustering[J]. Journal of Machine Learning Research, 2001, 2(12): 125-137.

[19] ALZATE C, SUYKENS J A K. Sparse kernel spectral clustering models for large-scale data analysis[J]. Neurocomputing, 2011, 74(9): 1382-1390.

[20] 张莉, 周伟达, 焦李成. 核聚类算法[J]. 计算机学报, 2002(6): 587-590.
ZHANG Li, ZHOU Weida, JIAO Licheng. Kernel clustering algorithm[J]. Chinese Journal of Computers, 2002(6): 587-590.

[21] TZORTZIS G, LIKAS A. The global kernel *K*-means clustering algorithm [C]//Proceedings of 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, China: IEEE, 2008: 1977-1984.

[22] LIU Y. Refined learning bounds for kernel and approximate *K*-means [J]. Advances in Neural Information Processing Systems, 2021, 34: 6142-6154.

[23] HUANG H C, CHUANG Y Y, CHEN C S. Multiple kernel fuzzy clustering[J]. IEEE Transactions on Fuzzy Systems, 2011, 20(1): 120-134.

[24] YU S, TRANCHEVENT L, LIU X, et al. Optimized data fusion for kernel *K*-means clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(5): 1031-1039.

[25] XIA R, PAN Y, DU L, et al. Robust multi-view spectral clustering via low-rank and sparse decomposition [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Quebec, Canada: AAAI, 2014: 2149-2155.

[26] GUO D, ZHANG J, LIU X, et al. Multiple kernel learning based multi-view spectral clustering[C]// Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014: 3774-3779.

[27] CHO Y, SAUL L. Kernel methods for deep learning[J]. Advances in Neural Information Processing Systems, 2009, 22: 342-350.

[28] STROBL E V, VISWESWARAN S. Deep multiple kernel learning[C]// Proceedings of the 12th International Conference on Machine Learning and Applications. Miami, USA: IEEE, 2013: 414-417.

[29] REBAI I, BENAYED Y, MAHDI W. Deep multilayer multiple kernel learning[J]. Neural Computing and Applications, 2016, 27(8): 2305-2314.

[30] LIU Y, LIAO S, HOU Y. Learning kernels with upper bounds of leave-one-out error[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Birmingham, UK: ACM, 2011: 2205-2208.

[31] JIU M, SAHBI H. Nonlinear deep kernel learning for image annotation[J]. IEEE Transactions on Image Processing, 2017, 26(4): 1820-1832.

[32] CHEN X, PENG X, DUAN R, et al. Deep kernel learning method for SAR image target recognition [J]. Review of Scientific Instruments, 2017, 88(10): 104706.

[33] ERHAN D, MANZAGOL P A, BENGIO Y, et al. The difficulty of training deep architectures and the effect of unsupervised pre-training[C]//Proceedings of the 2009 Artificial Intelligence and Statistics. Florida, USA: JMLR, 2009: 153-160.

(下转第 18 页)